

Information Geometry and Statistical Pattern Recognition

Shinto Eguchi

Institute of Statistics Mathematics
and Graduate University of Advanced Studies

Abstract

This paper discusses a geometry associated with U -divergence including ideas of U -models, U -loss functions of two versions. On the basis of the geometry we observe that U -divergence projection of a data distribution p onto U -model M_U associates with the Pythagorean relation for the triangle connection of p , q and q^* , for any q of the U -model where q^* denotes the point of M_U projected from p . This geometric consideration is implemented on the problem of statistical pattern recognition. U -Boost algorithm proposed in the practical application is shown to pursue iteratively the U -divergence projection onto U -model evolving by one dimension according to one iteration. In particular U -Boost algorithm released from the probability constraint reveals a novel property of statistical property beyond the notion of Fisher consistency, which helps us to understand the statistical meaning of AdaBoost.

Key words AdaBoost, exponential family, logistic model, maximum likelihood, U -divergence, U -model, U -loss function

1 Introduction

Information geometry has been growing a mathematical method for applying a variety of sciences crossing over the statistical science, information science, quantum physics, artificial intelligence since Amari [1] pioneered the important possibilities and foundations. In this article we discuss an application of information geometry focusing on a boosting algorithm, which is recently proposed as a novel algorithm for statistical pattern recognition in the machine learning community, [20], [26].

Which definition is information geometry? It would be difficult to answer this question in a word, however, it could be mildly said that this is an approach to discuss a parametric model of probability distributions, say $M = \{p_\theta(x) : \theta \in \Theta\}$ as a geometric object. In statistical theory a problem of inference is expressed not by M by the parameter space Θ . For example, for a point estimation a data distribution is assumed to be in M , however, it is often said that there exists a true value θ in the parameter space Θ . This tells us that the fundamental statement for statistical inference originally uses a relation between a manifold M and a coordinate space Θ , which is fundamental in differential geometry.

However, any close interaction could not be build between geometry and statistics in the early period of 20th century except for a suggestion that the model M is a Riemannian manifold with the information metric g , cf. Rao [21]. The information metric is defined by Fisher information matrix for any θ , of which the inverse gives the bound of variance matrices of unbiased estimators. This is proved by simply applying the Cauchy-Schwartz inequality, and it is formally equivalent to the uncertainty principle by Heisenberg. It is shown that the Riemannian connection $\bar{\nabla}$ associated with the information metric g does not play an intrinsic role in statistical theory, and hence a pair of linear connections, called e -connection $\nabla^{(e)}$ and m -connection $\nabla^{(m)}$, is formulated.

The reason to be formulated the connection pair naturally comes from a fact that a pair of statistical model and inference is necessary in statistical problem. Thus any single connection insufficiently tells the interactive relation between model and inference, while the connections $\nabla^{(e)}$ and $\nabla^{(m)}$ offer the optimal structures of model and inference, respectively. In other words different structures of optimality need different scales of flatness. Hence this dualistic formulation succeeds in understanding the notion of statistical sufficiency, exponential family, Fisher's maximum likelihood principle. It is interesting to see a fact that $\bar{\nabla} = \frac{1}{2}(\nabla^{(e)} + \nabla^{(m)})$ when we consider a dualistic relation of statistical model and statistical inference.

Let us look at a family of Gaussian distributions over \mathbb{R}^d with the density form

$$\mathcal{G}_d = \left\{ p_\theta(x) = (2\pi)^{-\frac{1}{2}d} \exp \left\{ -\frac{1}{2} \|x - \theta\|^2 \right\} : \theta \in \mathbb{R}^d \right\}, \quad (1)$$

where the variance matrix is fixed as the identity matrix.

In practice, \mathcal{G}_d is an elementary object in various subjects in Mathematics and we will see that \mathcal{G}_d is the most basic object in the information geometry as follows. In general the Kullback-Leibler (KL) divergence is defined by

$$\text{KL}(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

When we restrict the definition domain of to the KL divergence \mathcal{G}_d , we get that

$$\text{KL}(p_{\theta_1}, p_{\theta_2}) = \frac{1}{2} \|\theta_1 - \theta_2\|^2, \quad (2)$$

which is exactly proportional to the squared Euclidian distance. Thus the Gaussian distribution with mean vector θ and the identity variance matrix can be viewed as a point of the Euclidian space \mathbb{R}^d in the metric space. In general $\nabla^{(e)}$ and $\nabla^{(m)}$ are different; in \mathcal{G}_d they are the same as $\bar{\nabla}$ in which any geodesic line is nothing but the straight line. In the light of information geometry \mathcal{G}_d is Euclidean flat, which will be shown in a subsequent discussion. We result a fact that \mathcal{G}_d reduces to the Euclidean space, that is, the most important model in statistics can be reduced to the simplest geometry. This is closely related with a fact that Gauss's least square method is realized by linear projection in the Euclid space. Accordingly the pair of linear subspace and the projection onto it in E^d forms the optimal pair of statistical model and inference.

This relation can be extended more general situations beyond Gaussianity. We pay attention to a close relation of KL divergence and exponential family. It is shown from discussion of information geometry that an exponential model is dually flat with respect to e -connection and m -connection, and that KL divergence is expressed by the potential function and conjugate function over the exponential model, cf. [2]. In fact, we consider an exponential model defined by

$$\mathcal{E} = \{p_\theta(x) = \exp\{\theta^T b(x) - \varphi(\theta)\} : \theta \in \Theta\}.$$

where $\varphi(\theta)$ is the normalizing factor

$$\varphi(\theta) = \log \left[\int \exp\{\theta^T b(x)\} dx \right], \quad (3)$$

which is called cumulant function of $b(x)$. The cumulant function $\varphi(\theta)$ is a convex function on a convex set $\{\theta : \int \exp\{\theta^T b(x)\} dx < \infty\}$. The conjugate function is defined by

$$\varphi^*(\eta) = \sup_{\theta \in \Theta} \{\theta^T \eta - \varphi(\theta)\}.$$

The parameter transform from θ to η is given by

$$\eta = \frac{\partial \varphi(\theta)}{\partial \theta} = \int b(x) p_\theta(x) dx, \quad (4)$$

which is called Legendre transform. In accordance θ and η are affine parameters with respect to e -connection and m -connection, and that KL-divergence is expressed on the exponential model by these two coordinates as

$$\text{KL}(p_{\theta_1}, p_{\theta_2}) = \varphi^*(\eta_1) + \varphi(\theta_2) - \eta_1^T \theta_2,$$

where η_1 denotes the point coordinate mapped from θ_1 by the transformation (4). This conjugate convexity is a useful notion, which is commonly shared with a variety of fields in mathematical science. The fundamental notion of statistical sufficiency and efficiency can be understood by the use of this conjugate relation, [3].

In this paper we pursue the role of the functions of exp and log defining the cumulant function (3). Our main result will offer a variety of conjugate convexity beyond the cumulant function. We employ a real-valued function U in place of the function exp in (3) in which U is assumed to be convex with non-negative derivative u . For this objective we consider a set of non-negative functions without probability requirement. We will give a formulation that the function U naturally generates U -divergence, which associates with a dual linear connections. In the framework we consider a dually flat model, which will be extended the optimality result of the maximum likelihood under the exponential model. We apply this geometric framework to a problem in statistical pattern recognition. [20], [26].

Pattern recognition aims to decide the most plausible class-label of an object based on the feature vector. Statistical pattern recognition is a procedure to get a good pattern recognition by fully learning a training dataset, cf. [4], [18] for extensive discussion. It is reported that a biological brain system works a highly organized function for statistical pattern recognition. For example one can often recognize a person who he or she has not been met for a long time. This type of pattern recognition suggests that a highly integrated function is organized in a human brain system. A professional player of chess can neglect a vast of possible moves in the early stage, and simultaneously envisages the final stage. A professional potter can imagine the delicate color and texture after kilning even in the front of a woodturn table. An experienced wife recognizes all the psychological status of her husband from a bit of hid words or gestures. These remarkable performance is far beyond even the state-of-the-art performance of patter recognition machine.

Let us give a mathematical framework for problems of statistical pattern recognition [20]. Recently boosting algorithms are actively proposed and discussed in the field of artificial intelligence, [23], [10], [24], [13], [11]. In boosting algorithms AdaBoost has advantageous points over conventional methods and becomes popular as a universal procedure of statistical recognition. The learning algorithm is impressive in the respect in which the weight error rates are organized by drastic changes in a process of learning. The weighted error is updated into such that the best selected classification machine in the present step becomes the worst. In other

words, only the examples that the best machine fails to predict the class-label are featured up and the next best machine with respect the updated error rate is totally different from the present best machine.

In this paper we show that U -boost algorithm proposed by minimization of U -divergence inherits the geometry associated with U -divergence. We will see that this boost learning is interestingly understood from information geometric point of view. The structure of algorithm can be reduced to by Pythagorean theorem, which was proved in the era of ancient Greece, [20].

The rest of the paper is organized as follows. Section 2 introduces U -divergence and the associated flat model called U -model. Based on this framework the information-geometric discussion is developed in the comparison with the KL divergence and the maximum likelihood estimation. Section 3 gives a through application to statistical pattern recognition. Two version of U -boost algorithms are proposed and the statistical properties are investigated. In particular Eta-Boost algorithm is focused on the light of robustness. Section 4 summaries the concluding remarks and future problems.

2 U -divergence and U -model

2.1 U -divergence

Let Λ be a σ -finite measure on a data space \mathcal{Z} . We denote a space of all the non-negative functions with finite mass on the space \mathcal{Z} by

$$\mathcal{M} = \left\{ \mu : \mu(z) \geq 0 \text{ (a.e. } z \in \mathcal{Z}), \int_{\mathcal{Z}} \mu(z) d\Lambda(z) < \infty \right\},$$

and the subset with mass v by

$$\mathcal{M}_v = \left\{ \mu \in \mathcal{M} : \int_{\mathcal{Z}} \mu(z) d\Lambda(z) = v \right\}.$$

Throughout this paper our discussion can be applied for a case that \mathcal{Z} is metrizable, but in practice, it is sufficient for \mathcal{Z} to be a Euclid space or a discrete set. In a subsequent discussion we will suppose for the discussion on pattern recognition that \mathcal{Z} is essentially a set of finite numbers.

The space \mathcal{M} can be viewed as the space of Radon-Nicodim derivatives of finite measure dominated by Λ . In particular, $\mathcal{M}_{v=1}$ is the space of probability measures dominated by Λ . We say that a non-negative functional D defined over $\mathcal{M} \times \mathcal{M}$ is a contrast or divergence measure if D satisfies the first axiom of distance

$$D(\mu, \nu) = 0 \iff \mu = \nu \text{ (a.e. } \Lambda) \tag{5}$$

cf. [7]. Define a divergence over the function space \mathcal{M} using a real-valued function $U(t)$ as follows. We first assume that $U(t)$ is a convex function with non-negative derivative function $u(t) = U'(t)$ and let $\xi(u)$ be the inverse function of $u(t)$. Then we define as

$$D_U(\mu, \nu) = \int_{\mathcal{Z}} \left[U(\xi(\nu(z))) - U(\xi(\mu(z))) - \mu(z)\{\xi(\nu(z)) - \xi(\mu(z))\} \right] d\Lambda(z) \quad (6)$$

which we call U -divergence. We note that the convexity assumption of U leads D_U to satisfying the non-negativity and the requirement (5). In practice, it would be possible that any one-to-one transformation φ yields another divergence by

$$\int_{\mathcal{Z}} \left[U(\varphi(\nu(z))) - U(\varphi(\mu(z))) - u(\varphi(\mu(z)))\{\varphi(\nu(z)) - \varphi(\mu(z))\} \right] d\Lambda(z). \quad (7)$$

In this way U -divergence is a special choice by $\varphi = \xi$ in the form (7), which will be seen the specially useful properties.

Example 1. We first look at the most typical example which is defined by $U(t) = \exp(t)$. This implies that $u(t) = \exp(t)$, $\xi(u) = \log(u)$, which leads that U -divergence is nothing but KL divergence:

$$\text{KL}(\mu, \nu) = \int_{\mathcal{Z}} \left[\nu(z) - \mu(z) - \mu(z)\{\log(\nu(z)) - \log(\mu(z))\} \right] d\Lambda(z). \quad (8)$$

In the ordinary definition of KL divergence the first two terms in the right side of (8) cancel out because of the restriction to $\mathcal{M}_{v=1}$.

Example 2. We secondly consider

$$U(t) = \frac{1}{\beta + 1} (1 + \beta t)^{\frac{\beta+1}{\beta}}$$

and hence $u(t) = (1 + \beta t)^{1/\beta}$, $\xi(u) = (u^\beta - 1)/\beta$ and $U(\xi(u)) = u^{\beta+1}/(\beta + 1)$. Thus the corresponding divergence is

$$D_\beta(\mu, \nu) = \int_{\mathcal{Z}} \left[\frac{\{\nu(z)\}^{\beta+1} - \{\mu(z)\}^{\beta+1}}{\beta + 1} - \frac{\mu(z)[\{\nu(z)\}^\beta - \{\mu(z)\}^\beta]}{\beta} \right] d\Lambda(z), \quad (9)$$

which is called beta-divergence. We note that $\lim_{\beta \downarrow 0} D_\beta = \text{KL}$, which implies that if $\beta = 0$, beta-divergence reduces to KL-divergence. In the case of $\beta = 1$ it reduces to $\frac{1}{2} \int \{\mu(z) - \nu(z)\}^2 d\Lambda(z)$, or half of squared L_2 norm, cf. [25]. In a statistical framework this divergence is proposed to apply to principal component analysis, independent component analysis, cluster analysis from a robustness point of view. [14], [15], [19], [12].

Example 3. Thirdly we suggest a generic function

$$U(t) = (1 - \eta) \exp(t) + \eta t,$$

from which the Eta-divergence generated will be used for statistical pattern recognition, see Subsection 3.6.

2.2 Parametric model and geometry

We consider a model specified by a finite number of parameters in the function space \mathcal{M} as

$$M = \{\mu(z, \theta) : \theta \in \Theta\}. \quad (10)$$

Henceforth we assume that M is well-defined as a d -dimensional differentiable manifold with the coordinate system $\theta = (\theta^1, \dots, \theta^d)$ and the coordinate space Θ . For this smoothness assumption, the model function $\mu(z, \theta)$ is implicitly made under the integral sign $\int_{\mathcal{Z}} \cdot d\Lambda(z)$.

Let us consider a restriction of the definition domain of D to $M \times M$. Then the restriction

$$D(\theta_1, \theta_2) = D(\mu(\cdot, \theta_1), \mu(\cdot, \theta_2))$$

associates with a Riemannian metric $g^{(D)}$, a pair of linear connections $\nabla^{(D)}$ and $*\nabla^{(D)}$ on the differentiable manifold M as follows. We define

$$g^{(D)}(X, Y)(\mu) = -D(X|Y)(\mu)$$

for any vector fields X, Y on M , where the symbol $D(X|Y)$ denotes the following convention as, in general

$$D(X_1, \dots, X_n | Y_1, \dots, Y_m)(\mu) = X_1(\mu_1) \cdots X_n(\mu_1) Y_1(\mu_2) \cdots Y_m(\mu_2) D(\mu_1, \mu_2) \Big|_{\mu_1=\mu, \mu_2=\mu}.$$

By definition D has a minimum 0 on the diagonal $\{(\mu, \mu) : \mu \in M\}$, and the metric $g^{(D)}$ gives the primary approximation around the diagonal. Next we introduce $\nabla_X^{(D)} Y$ and $*\nabla_X^{(D)} Y$ as

$$g^{(D)}(\nabla_X^{(D)} Y, Z) = -D(XY|Z),$$

$$g^{(D)}(*\nabla_X^{(D)} Y, Z) = -D(Z|XY)$$

for any vector field Z . We note that two connections $\nabla^{(D)}$ and $*\nabla^{(D)}$ are both uniquely defined because of non-degeneracy of the metric $g^{(D)}$. The definition of $\nabla^{(D)}$ and $*\nabla^{(D)}$ is independent of $g^{(D)}$. However we find a close relation

$$\bar{\nabla}^{(D)} = \frac{1}{2}(\nabla^{(D)} + *\nabla^{(D)})$$

where $\bar{\nabla}^{(D)}$ is the metric connection with respect to $g^{(D)}$, cf. [6], [7]. In fact by definition we get that

$$Xg^{(D)}(Y, Z) = \frac{1}{2}X\{-D(Y|Z) - D(Z|Y)\} = g^{(D)}(\bar{\nabla}_X^{(D)} Y, Z) + g^{(D)}(Y, \bar{\nabla}_X^{(D)} Z)$$

which implies that $\bar{\nabla}^{(D)}$ is metric with respect to $g^{(D)}$. Next

$$g^{(D)}(\nabla_X^{(D)} Y - \nabla_Y^{(D)} X, Z) = -D(XY - YX|Z) = g^{(D)}([X, Y], Z),$$

which implies torsion-freeness of $\nabla^{(D)}$. Similarly we find the property for $*\nabla^{(D)}$. Hence $\bar{\nabla}^{(D)}$ is torsion-free and metric, which concludes that $\bar{\nabla}^{(D)}$ is the Riemannian connection because of the unique existence. In this sense $\nabla^{(D)}$ and $*\nabla^{(D)}$ are said to be conjugate.

In Section 1 we saw that the KL-divergence on the Gaussian model \mathcal{G}_d is expressed as (2), which implies that the associated geometry reduces to Euclidian on account of the above discussion.

Apply the general formula to U -divergence D_U defined in (6). The three geometric objects $g^{(U)}, \nabla^{(U)}, *\nabla^{(U)}$ associated with D_U are derived as follows. Let $(\partial/\partial\theta^1, \dots, \partial/\partial\theta^d)$ be the standard frame with respect to the coordinates $(\theta^1, \dots, \theta^d)$. Then

$$g^{(U)}\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right)(\theta) = \int_{\mathcal{Z}} \frac{\partial}{\partial\theta^i} \mu(z, \theta) \frac{\partial}{\partial\theta^j} \xi(\mu(z, \theta)) d\Lambda(z), \quad (11)$$

$$g^{(U)}\left(\nabla^{(U)} \frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^k}\right)(\theta) = \int_{\mathcal{Z}} \frac{\partial^2}{\partial\theta^i \partial\theta^j} \mu(z, \theta) \frac{\partial}{\partial\theta^k} \xi(\mu(z, \theta)) d\Lambda(z),$$

$$g^{(U)}\left(*\nabla^{(U)} \frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^k}\right)(\theta) = \int_{\mathcal{Z}} \frac{\partial}{\partial\theta^k} \mu(z, \theta) \frac{\partial^2}{\partial\theta^i \partial\theta^j} \xi(\mu(z, \theta)) d\Lambda(z).$$

We remark that all the three geometric objects depend only on ξ , where ξ is the inverse function of the derivative of U . KL-divergence is generated by $U = \exp$ with $\xi(u) = \log(u)$, which leads that

$$(g^{(\exp)}, \nabla^{(\exp)}, *\nabla^{(\exp)}) = (g, \nabla^{(m)}, \nabla^{(e)})$$

with the information metric g , m -connection $\nabla^{(m)}$ and e -connection $\nabla^{(e)}$.

If a model M is embedded as a flat model in \mathcal{M} , then M is also $\nabla^{(U)}$ -flat for any U -divergence. If M is embedded in $\{\xi(\mu) : \mu \in M\}$ by ξ -transform and the embedded form is flat, then M is $*\nabla^{(U)}$ -flat. Noting this fact we will build U -model in the following subsection. The most noteworthy about ever U -divergence is commonly that

$$\nabla^{(U)} = \nabla^{(m)}. \quad (12)$$

On the other hand, $*\nabla^{(U)}$ equals the e -connection $\nabla^{(e)}$ if and only if $U = \exp$.

This structure associated with U -divergence is contrast with that with f -divergence

$$D_f(\mu, \nu) = \int_{\mathcal{Z}} \left[f\left(\frac{\nu(z)}{\mu(z)}\right) \mu(z) - f'(1) \{\nu(z) - \mu(z)\} \right] d\Lambda(z),$$

where f is a convex function with $f(1) = 0$. The class of f -divergences and that of U -divergences intersects only at KL-divergence. That is to say, $f_0(t) = -\log(t)$ and $U_0(t) = \exp(t)$ generate the common divergence D_{KL} , while $D_f \neq D_U$ for any $f \neq f_0$ and any $U \neq U_0$. The characteristic associated with f -divergence is

$$g^{(f)} = g,$$

while that with U -divergence is the common property (12).

2.3 U -model

We introduce a specific linear model in the total space \mathcal{M} with reference to U -divergence. It is known that KL-divergence naturally associates with an exponential model, and that the sufficient statistic is available under the assumption. We explore how is the elegant property extended to U -divergence and the related model.

Let a generic function U be arbitrarily fixed. Then we define a kind of linear model using the derivative u of U :

$$M_U = \{\mu(z, \theta) = u(\theta^T b(z)) : \theta \in \Theta\}, \quad (13)$$

where $b(z)$ is assumed to be d -dimensional vector-valued function which is square-integrable with respect to the carrier measure Λ with no constant component in z . We will discuss this requirement later. Here the parameter space Θ is defined by

$$\Theta = \left\{ \theta \in \mathbb{R}^d : \int_{\mathcal{Z}} U(\theta^T b(z)) d\Lambda(z) < \infty \right\}.$$

Hence the convexity assumption of U yields that Θ is a convex set and that

$$\varphi_U(\theta) = \int_{\mathcal{Z}} U(\theta^T b(z)) d\Lambda(z) \quad (14)$$

is a convex function in θ , which is called the potential function. The potential function φ_U associates with the conjugate function

$$\varphi_U^*(\eta) = \sup_{\theta \in \Theta} \{\eta^T \theta - \varphi_U(\theta)\}$$

by the Fenchel duality. We can view $\eta = (\eta^1, \dots, \eta^d)$ as another coordinate system of M_U with the coordinate transformation from $\theta = (\theta^1, \dots, \theta^d)$ to $\eta = (\eta^1, \dots, \eta^d)$ defined by

$$\eta = \frac{\partial}{\partial \theta} \varphi_U(\theta), \quad \theta = \frac{\partial}{\partial \eta} \varphi_U^*(\eta). \quad (15)$$

In particular we get the explicit form of the transformation as

$$\eta = \int_{\mathcal{Z}} b(z) u(\theta^T b(z)) d\Lambda(z).$$

Thus U -model M_U has the canonical coordinate $\theta = (\theta^1, \dots, \theta^d)$ and the conjugate coordinate $\eta = (\eta^1, \dots, \eta^d)$, which are connected with the Legendre transform (15). By the use of dual coordinates U -divergence can be expressed as

$$D_U(\eta_1, \theta_2) = \varphi_U^*(\eta_1) + \varphi_U(\theta_2) - \eta_1^T \theta_2$$

on U -model M_U . We note that

$$-\frac{\partial^2}{\partial \eta_1 \partial \theta_2} D_U(\eta_1, \theta_2) = \text{Id} \quad (\text{identity matrix}),$$

which implies that U -model M_U is dually flat in the sense of $\nabla^{(U)}$ and ${}^*\nabla^{(U)}$ with affine parameters θ and η . From this it follows that M_U is a Hessian manifold as we get that

$$g^{(U)}\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right)(\theta) = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\varphi_U(\theta) \quad (16)$$

from the formula (11) in θ . Similarly,

$$g^{(U)}\left(\frac{\partial}{\partial\eta^i}, \frac{\partial}{\partial\eta^j}\right)(\eta) = \frac{\partial^2}{\partial\eta^i\partial\eta^j}\varphi_U^*(\eta) \quad (17)$$

with respect to η , of which the matrix of size $d \times d$ is the inverse matrix of that of (16).

We next consider for U -model to be in the space of probability densities, that is to say, $\mathcal{M}_{v=1}$ with mass 1. Define a probability U -model by

$$\bar{M}_U = \{\bar{p}(z, \theta) = u(\theta^T b(z) - \kappa(\theta)) : \theta \in \Theta\},$$

where $\kappa(\theta)$ is the normalizing factor defined by

$$\int_{\mathcal{Z}} u(\theta^T b(z) - \kappa(\theta)) d\Lambda(z) = 1.$$

The potential function on the coordinate space Θ is defined by

$$\bar{\varphi}_U(\theta) = \int_{\mathcal{Z}} U(\theta^T b(z) - \kappa(\theta)) d\Lambda(z) + \kappa(\theta).$$

Hence we get an exact expression of U -divergence over the probability U -model \bar{M}_U using the potential function $\bar{\varphi}_U(\theta)$ and the conjugate function $\bar{\varphi}_U^*(\eta_1)$ as follows:

$$D_U(\eta_1, \theta_2) = \bar{\varphi}_U^*(\eta_1) + \bar{\varphi}_U(\theta_2) - \eta_1^T \theta_2.$$

By an argument similar to that on U -model, a probability U -model is also seen to be dually flat.

2.4 m -projection onto U -model

We now discuss projection onto U -model M_U from a viewpoint of dually flatness of M_U . Let μ be fixed to satisfy $\mu \in \mathcal{M} - M_U$. Then we explore the minimization problem

$$\min\{D_U(\mu, \nu) : \nu \in M_U\}. \quad (18)$$

In fact, the substitution of $\mu(z, \theta)$ defined in (13) into ν of (18) leads that

$$D_U(\mu, \mu(\cdot, \theta)) = \varphi_U(\theta) - \theta^T b$$

by neglecting constants in θ , which is convex in θ , where

$$b = \int_{\mathcal{Z}} b(z)\mu(z)d\Lambda(z). \quad (19)$$

Hence the minimizer θ^* uniquely exists such that

$$\frac{\partial}{\partial \theta} \varphi_U(\theta^*) = b.$$

Therefore a Pythagorean Theorem holds in the function space \mathcal{M} :

$$D_U(\mu, \nu) = D_U(\mu, \nu^*) + D_U(\nu^*, \nu)$$

for any $\nu \in M_U$, where $\nu^*(z) = \mu(z, \theta^*)$. We define a linear subspace \mathcal{M} by

$$\mathcal{F}(\mu) = \left\{ \lambda \in \mathcal{M} : \int_{\mathcal{Z}} b(z) \{ \lambda(z) - \mu(z) \} d\Lambda(z) = 0 \right\}.$$

Then we get that

$$D_U(\lambda, \nu) = D_U(\lambda, \nu^*) + D_U(\nu^*, \nu)$$

for any $\lambda \in \mathcal{F}$. Therefore,

$$\nu^* = \operatorname{argmin} \{ D_U(\lambda, \nu) : \lambda \in \mathcal{F}(\mu) \} = \operatorname{argmin} \{ D_U(\lambda, \nu) : \nu \in M_U \},$$

where argmin denotes the argument attaining the minimization.

By a argument similar to the above discussion we can think projection of p of $\mathcal{M}_{v=1}$ onto the probability U -model \bar{M}_U . Thus the minimizer $\bar{\theta}^*$ of $D_U(p, \bar{p}(\cdot, \theta))$ in θ satisfies that

$$\int_{\mathcal{Z}} b(z) \{ u(\theta^{*T} b(z) - \kappa(\theta)) - p(z) \} d\Lambda(z) = 0.$$

In accordance with this, we get the Pythagorean theorem over the space $\mathcal{M}_{v=1}$:

$$D_U(q, \bar{p}) = D_U(q, \bar{p}^*) + D_U(\bar{p}^*, \bar{p})$$

for any $q \in \bar{\mathcal{F}}(p)$ and any $\bar{p} \in \bar{M}_U$, where $\bar{p}^*(z) = \bar{p}(z, \theta^*)$ and

$$\bar{\mathcal{F}}(p) = \left\{ q \in \mathcal{M}_{v=1} : \int_{\mathcal{Z}} b(z) \{ q(z) - p(z) \} d\Lambda(z) = 0 \right\}.$$

2.5 U -loss function on U -model

Let us consider the usual framework of statistical estimation. Let z_1, \dots, z_n be a random sample from a distribution with density function $p(z)$. Then we define a shifted U -model by

$$\tilde{M}_U = \{ \tilde{\mu}(z, \theta) = u(\theta^T (b(z) - \bar{b})) : \theta \in \Theta \}, \quad (20)$$

where $\bar{b} = \int b(z)p(z)d\Lambda(z)$. Here we note that if any components of $b(z)$ in U -model M_U as defined in subsection 2.3, then the sifted model \tilde{M}_U degenerates from M_U by one dimension.

We now discuss U -divergence projection

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} D_U(p, \tilde{\mu}(\cdot, \theta))$$

as discussed m -projection in section 2.4. We define U -loss function on the shifted U -model by

$$L_U(\theta) = \int_{\mathcal{Z}} U(\theta^T(b(z) - \bar{b}))d\Lambda(z). \quad (21)$$

Thus, neglecting constant terms in θ we get that

$$L_U(\theta) = D_U(p, \tilde{\mu}(z, \theta)). \quad (22)$$

Then U -loss function satisfies that

$$L_U(\theta) \geq L_U(\theta^*).$$

Proof follows from

$$L_U(\theta) - L_U(\theta^*) = D_U(\tilde{\mu}(\cdot, \theta^*), \tilde{\mu}(\cdot, \theta))$$

which is nothing but the Pythagorean theorem as in (22).

The empirical loss function is defined by substituting

$$\bar{b}_{\text{emp}} = \frac{1}{n} \sum_{i=1}^n b(z_i)$$

into \bar{b}_{emp} of (21) as follows:

$$L_U^{\text{emp}}(\theta) = \int_{\mathcal{Z}} U(\theta^T(b(z) - \bar{b}_{\text{emp}}))d\Lambda(z). \quad (23)$$

The solution is

$$\tilde{\theta}_U = \underset{\theta \in \Theta}{\text{argsolve}}\{\mathbb{E}_{\theta}(b(z)) = \bar{b}_{\text{emp}}\}, \quad (24)$$

where argsolve denotes the argument satisfying the given equation and \mathbb{E}_{θ} is the expectation with respect to the probability density

$$\tilde{p}(z, \theta) = \frac{u(\theta^T(b(z) - \bar{b}_{\text{emp}}))}{\int u(\theta^T(b(z') - \bar{b}_{\text{emp}}))d\Lambda(z')}. \quad (25)$$

Hence we summarize these results in the following theorem.

Theorem 1. *Assume that a true density function $p(z)$ satisfies that*

$$p(z) = \tilde{p}(z, \theta).$$

Then the estimator $\tilde{\theta}_U$ is asymptotically consistent for θ .

We find an interesting interpretation for this theorem. According to the original definition of $\tilde{\theta}_U$ it means the projection of $p(z)$ onto \tilde{M}_U by minimization of the empirical counterpart

of U -divergence. However, Theorem 1 claims that the projection can be viewed as projection onto

$$\{\tilde{p}(z, \theta) : \theta \in \Theta\}$$

embedded in $\mathcal{M}_{v=1}$ rather than \tilde{M}_U in \mathcal{M} . This view is striking beyond the usual idea between the model and inference, enables us to making decision in the whole space \mathcal{M} without probability restriction.

On the other hand, we can define U -loss function on the probability U -model \bar{M}_U by

$$\bar{L}_U(\theta) = \int_{\mathcal{Z}} \left[U(\theta^T b(z) - \kappa(\theta)) + \{\kappa(\theta) - \theta^T b(z)\} p(z) \right] d\Lambda(z). \quad (26)$$

Similarly we get for $\bar{\theta}^*$ projected from p onto the model \bar{M}_U that

$$\bar{L}_U(\theta) - \bar{L}_U(\bar{\theta}^*) = D_U(\bar{p}(\cdot, \theta^*), \bar{p}(\cdot, \theta)).$$

The empirical loss based on the random sample is

$$\bar{L}_U^{\text{emp}}(\theta) = \int_{\mathcal{Z}} U(\theta^T b(z) - \kappa(\theta)) d\Lambda(z) + \kappa(\theta) - \theta^T \bar{b}_{\text{emp}}, \quad (27)$$

and thus the solution minimizing the loss is given by

$$\bar{\theta}_U = \underset{\theta \in \Theta}{\text{argsolve}} \{ \bar{\mathbb{E}}_{\theta}(b(z)) = \bar{b}_{\text{emp}} \}. \quad (28)$$

where $\bar{\mathbb{E}}_{\theta}$ denotes the expectation with respect to $\bar{p}(z, \theta)$.

Consequently we observe that both the estimators defined by (24) and (28) has one-to-one correspondance with the statistic \bar{b}_{emp} . Hence the two have the same information as that of \bar{b}_{emp} . On the other hand the maximum likelihood estimator $\hat{\theta}_U$ under the probability U -model \bar{M}_U is defined to be maximized the log likelihood, and satisfies that

$$\hat{\theta}_U = \underset{\theta \in \Theta}{\text{argsolve}} \left\{ \sum_{i=1}^n \frac{u'(\theta^T b(z_i) - \kappa(\theta))}{u(\theta^T b(z_i) - \kappa(\theta))} (b(z_i) - \frac{\partial \kappa(\theta)}{\partial \theta}) = 0 \right\}, \quad (29)$$

which implies that $\hat{\theta}_U$ is not a function of only \bar{b}_{emp} . In statistical asymptotics the estimators $\bar{\theta}_U$ and $\hat{\theta}_U$ are both consistent for θ under the assumption where the true density function equals $\bar{p}(z, \theta)$. However we note that the supposed density function $\tilde{p}(z, \theta)$ in Theorem 1 is in general different from the $\bar{p}(z, \theta)$. What happens if these density functions coincide?

Theorem 2. *Let three estimators $\tilde{\theta}_U$, $\bar{\theta}_U$ and $\hat{\theta}_U$ be defined by (24), (28) and (29). Then if and only if $U = \exp$,*

$$\tilde{\theta}_U = \bar{\theta}_U = \hat{\theta}_U.$$

Proof. We observe that if $U = \exp$

$$\tilde{p}(z, \theta) = \bar{p}(z, \theta) = \exp(\theta^T b(z) - \kappa(\theta)),$$

which implies that the estimating equations defined in (24) and (28) are equal. Hence $\tilde{\theta}_U = \bar{\theta}_U$. Similarly the likelihood equation in (29) becomes the same, which concludes the proof of ‘if part’. The reverse statement is direct from the characterization of exponential function.

We will utilize this discussion to elucidate the statistical property of U -boost method for statistical pattern recognition.

3 statistical pattern recognition and U -Boost

3.1 statistical pattern recognition

Let us introduce a general framework of statistical pattern recognition. Let x be a feature vector in the feature space \mathcal{X} of a p -dimensional Euclidean space and y the class-label of x in the label set \mathcal{Y} . The pattern recognition aims to find a good solution of specifying y given x . Thus the solution is equivalent to giving a map h from \mathcal{X} to \mathcal{Y} . In this context $h(x)$ is called a classifier, which is often defined by way of a discriminant function F on $\mathcal{X} \times \mathcal{Y}$ as follows:

$$h_F(x) = \operatorname{argmax}\{F(x, y) : y \in \mathcal{Y}\}. \quad (30)$$

We note that the correspondence $F \mapsto h_F$ is not in general one-to-one. For example the most simplified function

$$f(x, y) = I(h_F(x) = y)$$

satisfies that $h_f = h_F$, where $I(A)$ denotes the indicator function of A . Thus we naturally conceive an equivalence relation on the space of discriminant functions as

$$F \sim G \stackrel{\text{def}}{\iff} h_F = h_G \quad \text{on } \mathcal{X}. \quad (31)$$

On any coset $\mathcal{C}[F] = \{G : G \sim F\}$ of F the classifier is invariant.

Next we consider statistical framework of the pattern recognition. The goal is incarnate in the form of the classifier by the use of a given n tuple examples (training data)

$$E_n = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}. \quad (32)$$

Thus the classifier $y = h(x)$ depends on E_n , we call a statistical classifier. The final goal is to construct the optimal statistical classifier. For this we have to set out the probabilistic assumption for the example set E_n and the class of discriminant functions $F(x, y)$. To incorporate the discussion in section 3 into this situation we change notation \mathcal{Z} and z into $\mathcal{X} \times \mathcal{Y}$ and (x, y) with the feature vector x and class-label y .

3.2 Boosting method

Let us discuss a problem of statistical pattern recognition with a feature space \mathcal{X} and a class-label set \mathcal{Y} for a given example set E_n defined by (32) in the product set. Suppose that a family \mathcal{H} of statistical classifiers is applicable for the problem. Then how can one organize the family to make a strong single classifier? Basically for any classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ of \mathcal{H} one can evaluate the empirical error rate

$$\text{Err}(h) = \frac{1}{n} \sum_{i=1}^n I(h(x_i) \neq y_i). \quad (33)$$

Hence as the best candidate in the family \mathcal{H} ,

$$h_{\text{naive}} = \text{argmin}\{\text{Err}(h) : h \in \mathcal{H}\} \quad (34)$$

would be selected. However there is room to be more carefully discussed the validity for this candidate h_{naive} . We cannot deny any possibilities in which there is another classifier h_* that extremely performs well only for the examples such that h_{naive} badly performs, namely only for

$$\{(\mathbf{x}_i, y_i) : h_{\text{naive}}(x_i) \neq y_i\}. \quad (35)$$

Thus we envisage a possible improvement of h_{naive} on performance using the complementary classifier h_* . In this sense it would be insufficient for us to adopt the classifier h_{naive} only.

In cognitive sciences it is researched that a biological brain organizes more rational rule for pattern recognition through the learning process. Recently boosting algorithms have been exploited to combine several individual classifiers with a slogan to analogy of brain in the community of machine learning.

The key idea in the boosting method is to embed any members h_1, \dots, h_d of \mathcal{H} into the space \mathcal{F} of discriminant functions as follows:

$$\mathcal{F} = \{F(x, y, \alpha) = \sum_{j=1}^d \alpha_j I(h_j(x) = y) : \alpha = (\alpha_1, \dots, \alpha_d) \in A\}. \quad (36)$$

We note that in this expression the classifier h_{naive} is expressed by all the null coefficients except for only one positive coefficient α_j in (36). If we rule out more reasonable way to give linear coefficients $\{\alpha_j\}$, we can construct a stronger classifier than h_{naive} . In fact a boosting method aims to offer a sequential algorithm

$$F(x, y, \alpha_1, \dots, \alpha_{t+1}) = F(x, y, \alpha_1, \dots, \alpha_t) + \alpha_{t+1} I(h_{t+1}(x) = y)$$

with an optimization design. The boosting method sequentially defines the optimal classifier h_{t+1} and coefficient α_{t+1} from the step t .

3.3 U -loss function for discriminant functions

Let $p(x, y) = P(y|x)q(x)$ be a probability distribution on the product space of a feature space \mathcal{X} and a class-label space \mathcal{Y} , where $P(y|x)$ is the conditional distribution of y given x and $q(x)$ is the marginal distribution. Assume that an example set $E_n = \{(x_i, y_i) : i = 1, \dots, n\}$ is an n -tuple realization from the distribution $p(y, x)$.

We write a combined discriminant function

$$F(x, y) = \alpha^T f(x, y)$$

using the embedding expression (36), where $f(x, y) = (I(h_1(x) = y), \dots, I(h_d(x) = y))$. Then we employ the shifted U -model

$$\tilde{\mu}_\alpha(y|x) = u(\alpha^T f(x, y) - b(x, \alpha)),$$

where $b(x, \alpha) = \sum_{y' \in \mathcal{Y}} \alpha^T f(x, y') p(y'|x)$. This is just an application of the general model discussed in subsection 2.5 along the context of statistical pattern recognition. Hence U -loss function is derived

$$L_U(\alpha) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} U(\alpha^T f(x, y) - b(x, \alpha)) q(x) dx$$

from the general definition of (21). By an argument similar to that in 2.5 we get that

$$L_U(\alpha) - L_U(\alpha^*) = D_U(\tilde{\mu}_{\alpha^*}, \tilde{\mu}_\alpha),$$

where

$$\alpha^* = \operatorname{argmin}_{\alpha \in A} D_U(p, \tilde{\mu}_\alpha).$$

The empirical U -loss function

$$L_U^{\text{emp}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \sum_{y \in \mathcal{Y}} U(\alpha^T \{f(x_i, y) - f(x_i, y_i)\}) \quad (37)$$

for a set $E_n = \{(x_i, y_i) : i = 1, \dots, n\}$ of examples, where

$$b(x_i, \alpha) = \alpha^T f(x_i, y_i).$$

On the other hand, we give a case with probability constraint, namely, a probability U -model

$$\bar{\mu}_\alpha(y|x) = u(\alpha^T f(x, y) - \kappa(x, \alpha)), \quad (38)$$

where $\kappa(\alpha)$ is a normalizing constant

$$\sum_{y \in \mathcal{Y}} u(\alpha^T f(x, y) - \kappa(x, \alpha)) = 1.$$

In accordance with this framework U -loss function is

$$\bar{L}_U(\alpha) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \left[U(\alpha^T f(x, y) - \kappa(x, \alpha)) - P(y|x) \{ \alpha^T f(x, y) - \kappa(x, \alpha) \} \right] q(x) dx.$$

Similarly the Pythagorean realtion

$$\bar{L}_U(\alpha) - \bar{L}_U(\alpha^*) = D_U(\bar{\mu}_{\alpha^*}, \bar{\mu}_{\alpha})$$

holds. The empirical U -loss function

$$\bar{L}_U^{\text{emp}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{y \in \mathcal{Y}} U(\alpha^T f(x_i, y) - \kappa(x_i, \alpha)) + \kappa(x_i, \alpha) - \alpha^T f(x_i, y_i) \right\} \quad (39)$$

is given.

Let us look at the most typical example of $U = \exp$ and then the two loss functions are

$$L_{\text{exp}}^{\text{emp}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \sum_{y \in \mathcal{Y}} \exp\{\alpha^T \{f(x_i, y) - f(x_i, y_i)\}\},$$

$$\bar{L}_{\text{exp}}^{\text{emp}}(\alpha) = -\frac{1}{n} \sum_{i=1}^n \log \left[\frac{\exp\{\alpha^T f(x_i, y_i)\}}{\sum_{y \in \mathcal{Y}} \exp\{\alpha^T f(x_i, y)\}} \right], \quad (40)$$

which are called the exponential loss and log loss functions, respectively. The two loss functions derive AdaBoost and LogitBoost. In statistical discussion the log loss is exactly minus of log-likelihood function for the logistic regression model, which is common in the community. On the other hand the exponential loss function is not know in the community, until it is proposed from the context of learning theory in the machine learning community [10].

Further it is recently elucidated in [16] that the typical loss functions are derived by KL-divergence as twin loss functions. We remark that the two empirical loss functions would be shown to generate consistent estimators when we apply Theorem 1 to the case of random sample. However, two different learning algorithms are proposed in the context of statistical pattern recognition.

3.4 U -Boost

We consider a sequential algorithm for minimization of the empirical loss functions introduced in 3.3. The basic idea is to a sequential projection of m -projection explored in subsection 2.4.

Assume that we have got an appropriate discriminant function $F(x, y)$ in the present step. Then we consider the best choice of a new classifier $h(x)$ to be combined with $F(x, y)$ by

$$F(x, y) \mapsto F^*(x, y) = F(x, y) + \alpha I(h(x) = y)$$

in the following update:

$$(\alpha^*, h^*) = \underset{(\alpha, h) \in \mathbb{R} \times \mathcal{H}}{\operatorname{argmin}} L_U^{\operatorname{emp}}(F(x, y) + \alpha I(h(x) = y)).$$

This minimization equivalently leads us the m -projection to construct a right triangle in the space \mathcal{F} satisfying the Pythagorean theorem.

The repetition of this operation generates the sequential minimization of U -loss function associated with the set of right triangles.

In practice, for a given example set E_n and family of classifiers \mathcal{H}_1 U -Boost algorithm for the version without probability constraint is proposed as follows:

A. We set $w_1(i, y) = \frac{1}{n(g-1)} I(y \neq y_i)$ as the initial weight distribution over E_n , where $g = \operatorname{card}(\mathcal{Y})$.

B. For a iteration number $t = 1, \dots, T$, the weighted error rate distribution is defined by

$$\epsilon_t(h) = \frac{1}{2} \sum_{i=1}^n \sum_{y \in \mathcal{Y}} w_t(i, y) I(y \neq y_i) \{f(x_i, y) - f(x_i, y_i) + 1\} \quad (41)$$

and then the following 3 sub-steps are executed

(B-1) Select $h_*^{(t)} = \underset{h \in \mathcal{H}_1}{\operatorname{argmin}} \epsilon_t(h)$

(B-2) Find $\alpha_t^* = \underset{\alpha}{\operatorname{argmin}} L_U^{\operatorname{emp}}(F_{t-1} + \alpha f_*^{(t)})$, where L_U^{emp} is defined by (37).

(B-3) Update $F_{t-1}(x, y)$ by $F_t(x, y) = F_{t-1}(x, y) + \alpha_t^* I(h_*^{(t)}(x) = y)$,
 $w_{t+1}(i, y) \propto u \{F_t(x_i, y) - F_t(x_i, y_i)\}$ updates the weighted error rate (41).

C. Finally, $h_{\text{final}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} F_T(x, y)$ is the classified to be completed, where

$$F_T(x, y) = \sum_{t=1}^T \alpha_t^* I(h_*^{(t)}(x) = y).$$

The U -Boost algorithm of version with probability constraint is given by replacing L_U^{emp} into $\bar{L}_U^{\operatorname{emp}}$ defined in (39) in the sub-step (B-2). Consequently U -Boost algorithm is a simple iterative algorithm to combine classifiers with different performance into the final classifier. The characteristic is focused on the dynamical changes of the weight distributions $w_t(i, y)$ over E_n for each t step. We will observe a remarkable property of the weight distribution which is common to all the U -Boost algorithms as follows: By definition the selected classifier $h_*^{(t)}$ minimizes the error rate with weighted by $w_t(i, y)$ while $h_*^{(t)}$ has the worst error rate $\frac{1}{2}$ according to the error rate $w_{t+1}(i, y)$ updated in the sub-step (B-2) by joining $h_*^{(t)}$ with coefficient α_t^* .

Theorem 3. *Every U -Boost algorithm satisfies that*

$$\epsilon_{t+1}(h_t) = \frac{1}{2}$$

for any step t .

Proof follows from a fact that the coefficient α_t^* in (B.2) has the gradient 0 of $L_U^{\text{emp}}(F_{t-1} + \alpha f_*^{(t)})$ with respect to α . See proof of Theorem 3 in [20] for detailed discussion.

Let us be back the case of $U(t) = \exp(t)$, and then we get the explicit solution in (B-2) as

$$\alpha_t^* = \frac{1}{2} \log \frac{1 - \epsilon_t(h_*^{(t)})}{\epsilon_t(h_*^{(t)})},$$

which is the counterpart of AdaBoost M2.

3.5 Equivalence with Bayes rule

In this section we discuss the statistical properties for U -Boost algorithm. Let a probability distribution on a direct space of a feature space \mathcal{X} and a class-label space \mathcal{Y} denote by

$$p(x, y) = P(y|x)q(x), \quad (42)$$

where $P(y|x)$ is the conditional distribution of y given x and $q(x)$ is the marginal distribution. Assume that an example set $E_n = \{(x_i, y_i) : i = 1, \dots, n\}$ is n -tuple realization from the distribution $p(y, x)$.

It is known that the Bayes rule

$$h_B(x) = \operatorname{argmax}\{P(y|x) : y \in \mathcal{Y}\} \quad (43)$$

provide the lower bound of the error rate under the assumption (43) if $P(y|x)$ is available.

In most statistical classifiers proposed are based on estimating the posterior distribution $P(y|x)$, [8], [9]. Which relation with the Bayes rule does U -Boost have? Noting that U -Boost is defined to minimize the empirical U -loss function we return the equivalence relation \sim naturally associated with statistical pattern recognition as explored in subsection 3.1. It is efficient to consider the coset

$$\mathcal{F}_B = \{F(x, y) : F \sim P\}.$$

We discuss the shifted U -model and probability U -modeling a nonparametric ways follows.

$$\mathcal{M}_U = \{\mu_F(y|x) = u(F(x, y) - b_F(x)) : F \in \mathcal{F}\},$$

$$\bar{\mathcal{M}}_U = \{\bar{\mu}_F(y|x) = u(F(x, y) - \kappa_F(x)) : F \in \mathcal{F}\}.$$

where $\kappa_F(x)$ is the normalizing factor and

$$b_F(x) = \sum_{y' \in \mathcal{Y}} F(x, y')p(y'|x). \quad (44)$$

Then we get the following theorem.

Theorem 4. Assume that there exists a discriminant function F^* in a certain class \mathcal{F} such that

$$u(F^*(x, y) - b_{F^*}(x)) = c(x)P(y|x), \quad (45)$$

where $c(x)$ is a positive function. Then,

$$F^* = \operatorname{argmin}\{L_U(F) : F \in \mathcal{F}\}, \quad (46)$$

where

$$L_U(F) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} U(F(x, y) - b_F(x))q(x)dx.$$

Proof. By definition,

$$\begin{aligned} & L_U(F) - L_U(F^*) - D_U(\mu_{F^*}, \mu_F) \\ &= - \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \mu_{F^*}(y|x) \{F^*(x, y) - b_{F^*}(x) - F(x, y) + b_F(x)\} q(x) dx. \end{aligned}$$

Further, from the assumption (45) it follows that

$$\begin{aligned} & L_U(F) - L_U(F^*) - D_U(\mu_{F^*}, \mu_F) \\ &= \int_{\mathcal{X}} c(x)q(x) \left[\sum_{y \in \mathcal{Y}} P(y|x) \{F^*(x, y) - b_{F^*}(x) - F(x, y) + b_F(x)\} \right] dx = 0. \quad (47) \end{aligned}$$

This is because the bracket term in the right side of (47) vanishes noting the definition of the shift term $b_F(x)$ as given in (44). Hence we conclude (46) from the property (5) of U -divergence.

The assumption (45) of Theorem 4 implies that $F^* \in \mathcal{F}_B$, or equivalently that F^* is equivalent to the Bayes rule. Accordingly the minimization of the abstract loss function $L_U(F)$ is consistent with the Bayes rule.

In practice the empirical loss function (37), namely the empirical expectation from the example set E_n , is only available. In this setting a parameter vector α of finite dimension is sequentially optimized by U -Boost algorithm. Under the assumption (43) the statement of Theorem 4 asymptotically holds for the size n of examples.

3.6 EtaBoost

We observe that U -Boost algorithm can be applied to a problem of statistical pattern recognition when the generic function U is fixed to satisfy the convexity with non-negative derivative u . We consider which U efficiently works for a specific problem among the class. We focus on robustness in statistical pattern recognition.

For this we precisely investigate the U -function given in Example 3 of subsection 2.1 as

$$U_\eta(t) = (1 - \eta) \exp(t) + \eta t,$$

where η is a constant with $0 < \eta < 1$. Hence

$$u_\eta(t) = (1 - \eta) \exp(t) + \eta, \quad \xi_\eta(u) = \log \frac{u - \eta}{1 - \eta},$$

which generate the divergence

$$D_\eta(\mu, \nu) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \left[\nu(x, y) - \mu(x, y) - \{\mu(x, y) - \eta\} \log \frac{\nu(x, y) - \eta}{\mu(x, y) - \eta} \right] q(x) dx,$$

which we call Eta-divergence. Hence, from the original definition (38) we see that the probability U -model for a discriminant function $F(x, y) = \alpha^T f(x, y)$ is

$$\bar{\mu}_\eta(y|x, \alpha) = (1 - \eta) \exp\{\alpha^T f(x, y) - \kappa(x, \alpha)\} + \eta,$$

where the normalizing function is given by

$$\kappa(x, \alpha) = \log \frac{1 - \eta}{1 - g\eta} + \log \left[\sum_{y' \in \mathcal{Y}} \exp\{\alpha^T f(x, y')\} \right]$$

with $g = \text{card}(\mathcal{Y})$. This is rewritten as

$$\bar{p}_\eta(y|x, \alpha) = \{1 - \eta(g - 1)\} P_L(y|x, \alpha) + \eta \sum_{y' \neq y} P_L(y|x, \alpha), \quad (48)$$

where

$$P_L(y|x, \alpha) = \frac{\exp\{\alpha^T f(x, y)\}}{\sum_{y' \in \mathcal{Y}} \exp\{\alpha^T f(x, y')\}}.$$

The probability model (48) provides the following interpretation. We make an ideal assumption such that the conditional distribution $P(y|x)$ given x is modeled by a logistic model $P_L(y|x, \alpha)$. However, we consider a practical situation in which the ideal assumption breaks down by some reason, and so that the class-label y is erroneously observed with probability η . Thus we observe that the concluding probability reduces to $\bar{p}_\eta(y|x, \alpha)$. In this sense Eta-divergence D_η associates with the generative model with mislabels.

We call the boost generated by U_η Eta-Boost. In accordance with the above interpretation Eta-Boost is a robust procedure. In practice, Eta-Boost of the probability version is equivalent to the method proposed by [5] for a binary regression analysis with noisy data, see [26] for detailed discussion.

When we release the probability constrains Eta-Boost associates with a model

$$\tilde{p}_\eta(y|x, \alpha) = \frac{(1 - \eta) \exp\{\alpha^T f(x, y) - b(x, \alpha)\} + \eta}{(1 - \eta) \sum_{y' \in \mathcal{Y}} \exp\{\alpha^T f(x, y') - b(x, \alpha)\} + g\eta}$$

from (25), which can be rewritten by

$$\tilde{p}_\eta(y|x, \alpha) = \{1 - \tilde{\eta}(x)(g - 1)\} P_L(y|x, \alpha) + \tilde{\eta}(x) \sum_{y' \neq y} P_L(y|x, \alpha),$$

where

$$\tilde{\eta}(x) = \frac{\eta}{(1 - \eta) \sum_{y' \in \mathcal{Y}} \exp\{\alpha^T f(x, y') - b(x, \alpha)\} + g\eta}.$$

In this way the probability of mislabel is given by $\tilde{\eta}(x)$ depending on x , which allocates higher probability as x is close to the decision boundary. It means that the Eta-Boost without the probability constraint adaptively performs in the sense that the class-label y is predicted with considerations of mislabel according to the distance of x to the decision boundary. This excellent property of Eta-Boost is pursued in the light of several real data analyses, cf. [26].

4 Conclusions and future perspective

We have presented a geometry associated with U -divergence including ideas of U -models, U -loss functions of two versions. This geometric consideration leads to a special application to statistical pattern recognition. U -Boost algorithm associates with iteratively the U -divergence projection onto U -model evolving by one dimension according to one iteration. U -Boost algorithm of the version without the probability constraint, typically AdaBoost is shown to perform the novel property of statistical property beyond the notion of Fisher consistency. We discuss the property invariant over the coset of the Bayes rule with respect to the equivalence relation a natural requirement of pattern recognition. In the research area of statistical learning theory the method of support vector machine has been developed parallel to the boosting method, cf. [27], [13]. Basically the two paradigms have different objectives, in which an approach is recently proposed to connect the two methods based on the idea of soft margin, cf. [22]. As a future project we mention an embedding of U -loss function to a kernel space, which is closely related with a problem of characterization of U -divergence class. In particular, it needs a study of infinite-dimensional analysis on U -model as done by the use of the theory of Orlicz space in [17].

Recently there appear a vast of data sets of higher dimension along rapidly growing research activities in the genome sciences. For example the micro-array technology enables to the simultaneous observations to gene expressions for a large group of genes. This information from the gene expression data should be related with difficult diseases, sensitivity for medication, and so that the problem is directly formulated as that of pattern recognition in which the feature vector is vector of gene expression and, for example, class-label denotes the occurrence of considerable drug sensitivity. However it is known that there is an unbalance relation between the number p and the sample size n , which leads to spurious discovery for the relation of the particular gene expression and disease. The problem is addressed as $n \ll p$, which motivates a variety of approaches. It is interesting that we find a solution of the problem in the class of U -Boost algorithms

Acknowledgments: I would like to express my thanks to Noboru Murata, Takafumi Kanamori and Takashi Takenouchi for joining the project related with this article and Hiromori Fujisawa and Masanori Kawakita for many suggestions and comments for improvement of the draft.

References

- [1] Amari, S. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.
- [2] Amari, S., and Nagaoka, H. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, Oxford, 2000.
- [3] Barndorff-Nielsen, O. E. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, 1978.
- [4] Bishop, C. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- [5] Copas, J. Binary Regression Models for Contaminated Data. *J. Royal Statist. Soc. B*, **50** (1988), 225-265.
- [6] Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann Statist.* **11** (1983), 793-803.
- [7] Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J.* **22** (1992), 631-647.
- [8] Eguchi, S., and Copas, J. B. Recent developments in discriminant analysis from an information geometric point of view. *J. Korean Statist. Soc.* **30** (2001), 247-264.
- [9] Eguchi, S., and Copas, J. B. A class of logistic type discriminant functions. *Biometrika* **89** (2002), 1-22.
- [10] Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences* **55** (1997), 119-139.
- [11] Friedman, J. H., Hastie, T., and Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** (2000), 337-407.
- [12] Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto T. and Matsuura M. Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics* **20** (2004), 718-726.
- [13] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning*. Springer, New York, 2001.
- [14] Higuchi, I. and Eguchi S. The influence function of principal component analysis by self-organizing rule. *Neural Computation* **10** (1998), 1435-1444.
- [15] Kamiya, H. and Eguchi, S. A class of robust principal component vectors. *J. Multivariate Analysis* **77** (2001), 239-269.

- [16] Lebanon, G., and Lafferty, J. Boosting and maximum likelihood for exponential models. *Advances in Neural Information Processing Systems* **14** (2002).
- [17] Pistone, P. and Sempì, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **23** (1995), 1543–1561.
- [18] McLachlan, G. J. *Discriminant analysis and statistical pattern recognition*. Wiley, New York, 1992.
- [19] Minami, M., and Eguchi, S. Robust blind source separation by beta-divergence. *Neural Computation* **14** (2002), 1859–1886.
- [20] Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. Information geometry of U-Boost and Bregman divergence. To appear in *Neural Computation* (2004).
- [21] Rao, C. R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** (1945), 81–91.
- [22] Rätsch, G., Onoda, T. and Müller K.-R. Soft Margins for AdaBoost. *Machine Learning* **42** (2001), 287–320.
- [23] Schapire, R. E. The strength of weak learnability. *Machine Learning* **5** (1990), 197–227.
- [24] Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.*, **26** (1998), 1651–1686.
- [25] Scott, D. W. Parametric statistical modeling by minimum integrated square error. *Technometrics* **43** (2001), 274–285.
- [26] Takenouchi, T., and Eguchi, S. Robustifying AdaBoost by adding the naive error rate. *Neural Computation* **16** (2004), 767–787.
- [27] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.